NON-NEGATIVE MATRIX FACTORIZATION BASED COMPENSATION OF MUSIC FOR AUTOMATIC SPEECH RECOGNITION

Bhiksha Raj[1,3]*, Tuomas Virtanen*[2]*, Sourish Chaudhuri*[3]*, Rita Singh*[3]

[1]Walt Disney Research, Pittsburgh PA, USA,
[2]Department of Signal Processing, Tampere University of Technology, Finland
[3]Carnegie Mellon University, Pittsburgh PA, USA,

bhiksha@cs.cmu.edu, tuomas.virtanen@tut.fi, sourishc@cs.cmu.edu, rsingh@cs.cmu.edu

ABSTRACT

This paper proposes to use non-negative matrix factorization based speech enhancement in robust automatic recognition of mixtures of speech and music. We represent magnitude spectra of noisy speech signals as the non-negative weighted linear combination of speech and noise spectral basis vectors, that are obtained from training corpora of speech and music. We use overcomplete dictionaries consisting of random exemplars of the training data. The method is tested on the Wall Street Journal large vocabulary speech corpus which is artificially corrupted with polyphonic music from the RWC music database. Various music styles and speech-to-music ratios are evaluated. The proposed methods are shown to produce a consistent, significant improvement on the recognition performance in the comparison with the baseline method. Audio demonstrations of the enhanced signals are available at http://www.cs.tut.fi/~tuomasv.

Index Terms: noise robustness, automatic speech recognition, non-negative matrix factorization, speech enhancement

1. INTRODUCTION

The problem of recognizing speech in the presence of non-stationary noises remains a difficult one without a satisfactory solution to date. A large number of algorithms have been proposed in the literature to address the more general problem of mitigating the effect of noise on the speech signal. Many of these attempt to reduce the noise from the speech signal itself [1, 2]. Other techniques *e.g.* [3, 4] modify features derived from the signal to reduce the effect of noise.

Many of the above mentioned methods are quite effective, often achieving dramatic improvements in recognition accuracy when speech is corrupted by stationary or slowly-varying noise. Unfortunately, the same improvements are not achieved when the noise is non-stationary. The reason for this is simple to intuit - these algorithms require (statistical) estimates of the spectral characteristics of the noise. Non-stationary noise changes quickly, often as quickly as the speech signal itself. Any local characteristics that one may estimate for the noise, based on past samples of speech, or even on a window of samples around the current sample, are unlikely to be representative of the noise affecting the current segment.

It can be reasoned, therefore, that techniques that can effectively estimate the noise at the current instant based on the current sample of the noisy speech are more likely to be effective when the noise is fast varying. However, since we have only the noisy speech to estimate the instantaneous noise from, we require stronger *a priori* information about the signals involved, namely the speech and the noise.

A number of techniques have been proposed based on this principle. These techniques generally attempt to model the temporal dynamics of either the speech [5], or the corrupting noise [6] or both [7, 8] by HMMs or linear dynamical systems, in order to aid localization of the current noise characteristics. However a simple characterization such as a linear dynamical system (or the more coarse Gaussian mixture model) is insufficiently detailed for signals such as music or speech, which have nearly unlimited range of variation. More detailed characterizations such as HMMs or graphical models [8] are only useful for very restricted noises as they are very detailed tend to overfit to the specific instances of noise they are trained from.

In this paper we follow a different approach. In previous work we (and other researchers) have demonstrated that non-negative spectral factorization methods, including those based on non-negative matrix factorization (NMF) [9] and latent-variable analysis (LVA) [10], can be effectively used for signal separation. These methods represent signals by a compositional model that characterizes their spectra as a weighted linear combination of additive units, or bases that combine to compose it. By appropriately learning these bases, it becomes possible to attempt to separate out mixtures of sounds. The mixed sound is modelled as a composition of the bases of all the contributing sources. Through the application of appropriate constraints, we can the estimate the contributions of the individual bases, and thereby reconstitute the individual sources contributing to the mixture.

In this paper we show that the NMF-based approach, described in Sections 2-3, is also capable of generating enhanced signals that significantly improve recognition on speech corrupted by a highly non-stationary signal, specifically music. Unlike the methods of [9, 10], we will use an exemplar-based method [11, 12] to learn overcomplete sets of bases for the signals. As in the previous techniques, both for NMF-based separation and statistical techniques such as [6, 7, 8], we require characterizations of both signals, *i.e.* speech and the corrupting signal (music in our case). However, we also demonstrate that the separation obtained using the exemplar based method generalizes to cases where the specific music type is not known. Experiments described in Section 4 on speech corrupted by music, a notoriously difficult non-stationary signal to compensate for, show that while large improvements in recognition accuracy can be obtained if the characterizations for both the speaker and the specific genre of music in the signal are known, the method is equally effective even if only generic characterizations that represent ensembles of music or speakers are available.

2. FEATURE REPRESENTATION

The basic feature representation we employ for NMF-based enhancement is the magnitude spectrogram. To obtain this, we compute series of short-time Fourier transforms (STFT) from hamming-windowed frames of the signal, and take the absolutive values of the resulting spectral vectors. For automatic speech recognition, Mel-frequency cepstral coefficients are computed from the enhanced magnitude spectrum representation.

The optimal analysis window length for NMF-based signal enhancement and speech recognition are different — for the former we have found an analysis window between 40ms and 64ms to be optimal, whereas speech recognition works best with an analysis window of about 25ms. As a result, we revert the NMF-enhanced magnitude spectrogram back to a time-domain signal, as described in Section 3.4. The reconstituted signal is used to compute features for recognition.

A key aspect of the magnitude spectrographic representation is that the magnitude spectrogram of the sum of two signals is approximately equal to the magnitude spectrogram of the individual signals. Let $y[n]$ be a noisy speech signal that is the sum of clean speech signals $s[n]$ and noise $m[n]$, n being the discrete time index. Let us denote the magnitude spectrum vectors of the signals in frame t as $\mathbf{y}_t$, $\mathbf{s}_t$, and $\mathbf{m}_t$, respectively. The sequence the spectral column vectors from the whole recording are grouped into matrices $\mathbf{Y}$, $\mathbf{S}$, and $\mathbf{M}$, respectively. Thus, we can approximate $\mathbf{Y} \approx \mathbf{S} + \mathbf{M}$.

3. SEPARATING SPEECH FROM NOISE

The principle behind NMF-based separation of speech from noise is this. Per the compositional model presented in Section 3.1, the spectral vector for clean speech is composed from the bases of speech, and the spectrum for the noise that corrupts the speech is composed from the bases for noise. The noisy speech, being an addition of speech and noise, is therefore a composition of the combined bases from speech and noise. The contributions of the individual bases to the noisy speech can be estimated as described in 3.3, segregating the contributions of speech bases from the noise bases, so that the speech in the mixture can be separated from the noise and synthesized as described in Section 3.4.

3.1. The Compositional Model

The compositional model represents the magnitude spectrum $\mathbf{s}_t$ of speech in frame t as a weighted linear non-negative combination of basis vectors $\mathbf{b}^{\mathrm{s}}_i$ as

$$\mathbf{s}_t = \sum_{i=1}^{S} \mathbf{b}^{\mathrm{s}}_i w^{\mathrm{s}}_{i,t} \tag{1}$$

where $\mathbf{b}^{\mathrm{s}}_i$ is the i^{th} speech basis vector and $w^{\mathrm{s}}_{i,t}$ is the weight of the basis in frame t, and S is the number of speech basis vectors.

If we represent the set of basis vectors using matrix $\mathbf{B}_{\mathrm{s}} = \left[\mathbf{b}^{\mathrm{s}}_1, \ldots, \mathbf{b}^{\mathrm{s}}_S\right]$, the model and the weights using matrix $[\mathbf{W}_{\mathrm{s}}]_{i,t} = w^{\mathrm{s}}_{i,t}$, we can write the model for the speech spectrogram as the product of matrices $\mathbf{B}_{\mathrm{s}}$ and $\mathbf{W}_{\mathrm{s}}$:

$$\mathbf{S} = \mathbf{B}_{\mathrm{s}}\mathbf{W}_{\mathrm{s}} \tag{2}$$

Similarly, the noise is modeled as the weighted sum of noise basis vectors $\mathbf{b}^{\mathrm{m}}_i$, $i = 1, \ldots, M$, where M is the number of noise basis vector. When the noise basis vectors are grouped into matrix $\mathbf{B}_{\mathrm{m}}$ and the noise weights into matrix $\mathbf{W}_{\mathrm{m}}$, the model for the noise spectrogram can be written as $\mathbf{M} = \mathbf{B}_{\mathrm{m}}\mathbf{W}_{\mathrm{m}}$.

The model for the noisy speech spectrogram $\mathbf{Y} \approx \mathbf{S} + \mathbf{M}$ can be written as

$$\mathbf{S} \approx \mathbf{B}\mathbf{W}, \tag{3}$$

where $\mathbf{B} = [\mathbf{B}_{\mathrm{s}}\mathbf{B}_{\mathrm{m}}]$ be a matrix that combines the bases for speech and noise into a single matrix, and $\mathbf{W} = [{\mathbf{W}_{\mathrm{s}}}^{\top}{\mathbf{W}_{\mathrm{m}}}^{\top}]^{\top}$ combines the weights into a single matrix. If there are S bases for speech (*i.e* $\mathbf{B}_{\mathrm{s}}$ is $D \times S$, where D is the dimensionality of the spectral vectors) and M bases for the noise ($\mathbf{B}_{\mathrm{m}}$ is $D \times M$), then the total number of bases in $\mathbf{B}$ is $S + M$, *i.e.* $\mathbf{B}$ is a $D \times (S + M)$ matrix. $\mathbf{W}$ is a $(S + M) \times T$ matrix.

All bases and weights in (3) are strictly non-negative. The intuition behind this is that any sound is composed by constructive composition of various components. For instance, a segment of music may be composed by additive composition of the notes that comprise it. Cancellation, which is a major component of any decomposition in terms of additive bases, rarely, if ever, factors into the composition of a sound, except by careful design. In fact, it has been found out that the non-negativity restriction alone is sufficient for even blind separation of sources [9].

3.2. The bases

The bases $\mathbf{b}^{\mathrm{s}}_i$ and $\mathbf{b}^{\mathrm{m}}_i$ which form the columns of $\mathbf{B}$ reside in the same domain as the data vectors $\mathbf{y}_t$, *i.e.* they too are spectral vectors and represent the spectral magnitudes of the composing signals. Equation 1 does not specify how the bases are obtained, and this remains a matter of choice. It is possible to obtain a data-driven estimate of an set of bases by analysis of example data. Two methods are immediately available - latent variable decompositions [10] and non-negative matrix factorization (NMF) [9]. However in this paper we have found an *exemplar-based* characterization [11, 12] to be most effective.

Exemplar-based characterizations use realizations of spectral vectors from the source signals itself as the bases. These bases may simply be drawn randomly from a collection of spectral vectors for the source. Thus each spectral vector is explained as being a linear combination of the exemplar vectors from the source. Although this defies any clear semantic interpretation (such as notes being the elementary units of music), such bases nevertheless have useful theoretical properties, particularly in the context of signal enhancement, as explained in [11].

We obtain the set of speech basis vectors $\mathbf{B}_{\mathrm{s}}$ as the magnitude of the DFT of randomly drawn frames from training examples of speech. Similarly, the set of noise basis vectors $\mathbf{B}_{\mathrm{m}}$ is obtained from from training examples of the corrupting signal. The detailed explanation of the data sets are described in Section 4.

3.3. Estimating Weights

Once a set of bases $\mathbf{B}$ is given, the weights with which they must be combined to optimally compose the spectral vectors in $\mathbf{Y}$ can be determined using either the EM algorithm from [10] or one of various NMF-based update rules [13]. In this paper we employ the NMF update rule that minimizes a generalized Kullback-Leibler divergence between the spectral vectors in $\mathbf{S}$ and the composition $\mathbf{BW}$. This rule estimates the weights through iterations of:

$$\mathbf{W} = \mathbf{W} \otimes \frac{\mathbf{B}^{\top}.[\frac{\mathbf{Y}}{\mathbf{B}.\mathbf{W}}]}{\mathbf{B}^{\top}.\mathbf{1}} \tag{4}$$

where $\mathbf{1}$ is a D-by-T matrix of ones. The operation $\otimes$ represents element-wise multiplication, and all divisions too are element wise. We initialize all the weights $\mathbf{W}$ to unity and apply the update 200 times. After that the weight matrices $\mathbf{W}_{\mathrm{s}}$ and $\mathbf{W}_{\mathrm{m}}$ for speech and noise, respectively, are obtained by splitting $\mathbf{W}$ as $\mathbf{W} = [\mathbf{W}_{\mathrm{s}}^{\top}\mathbf{W}_{\mathrm{m}}^{\top}]^{\top}$.

3.4. Signal reconstruction

The minimum-mean-squared-error estimate of $\mathbf{S}$, *i.e.* the contribution of speech to $\mathbf{Y}$ can be extracted as:

$$\mathbf{S} = \mathbf{Y} \otimes \frac{\mathbf{B}_{\mathrm{s}}\mathbf{W}_{\mathrm{s}}}{\mathbf{B}_{\mathrm{s}}\mathbf{W}_{\mathrm{s}} + \mathbf{B}_{\mathrm{m}}\mathbf{W}_{\mathrm{m}}} \tag{5}$$

The reconstituted speech spectrogram is then converted back to a time-domain signal by combining it with the phase obtained from the complex spectrogram of the noisy signal, applying an inverse STFT, and overlap-add combination of the frames. The time-domain signal is then further used for speech recognition. The above procedure can also be viewed as filtering the noisy signal with a time-varying filter defined by $(\mathbf{B}_{\mathrm{s}}\mathbf{W}_{\mathrm{s}})/(\mathbf{B}_{\mathrm{s}}\mathbf{W}_{\mathrm{s}} + \mathbf{B}_{\mathrm{m}}\mathbf{W}_{\mathrm{m}})$, similarly to Wiener filtering.

4. EXPERIMENTAL EVALUATION

4.1. Acoustic material and recognizer

We performed speech recognition experiments on digital mixtures of speech and music. The CMU-Sphinx HMM-based continuous-density speech recognizer was used for all experiments. Since NMF-based separation typically requires bases to characterize the speaker, we used a somewhat unconventional setup. Acoustic models with 1000 tied states, each modelled by a mixture of 8 Gaussians, were trained from the Resource Management database. As features we use MFCCs and their deltas and double-deltas calculated in 25 ms frames. Speakers from the *training* components of the Wall Street Journal were used as our test set. A total of 3775 utterances distributed approximately uniformly across 83 speakers were used as our test set. The remaining data from each speaker were used to train bases for the speaker where necessary.

The test utterances were corrupted by digital addition of music. For the music, we used the RWC database [14], which is a professionally produced polyphonic music database containing many different music styles. The included styles are "classical" from the RWC Classical database, and "jazz", "latin", and "world" from the RWC Genre database. Some of the recordings contain vocals. As speech may be confused with sung vocals, we simplify the recognition task by semi-automatically discarding music material containing singing, using simple rules to discard shorter segments, derived from MIDI references, and by listening to the rest. The first minute of each recording was segmented out and added to a collection to be used as "training data". Random segments from the remaining material were used for corrupting the speech. The above procedure resulted in total 339, 149, and 281 seconds of training material for the jazz, latin, and world main categories, respectively, and 1015, 368, and 674 seconds of testing material. For the classical music we had nearly 3 hours of test data and over 20 minutes of training data. All the material was downsampled from 44.1 to 16 kHz sampling frequency, and downmixed to mono. The test data were used to corrupt the speech and the training data were used to learn bases for the music types.

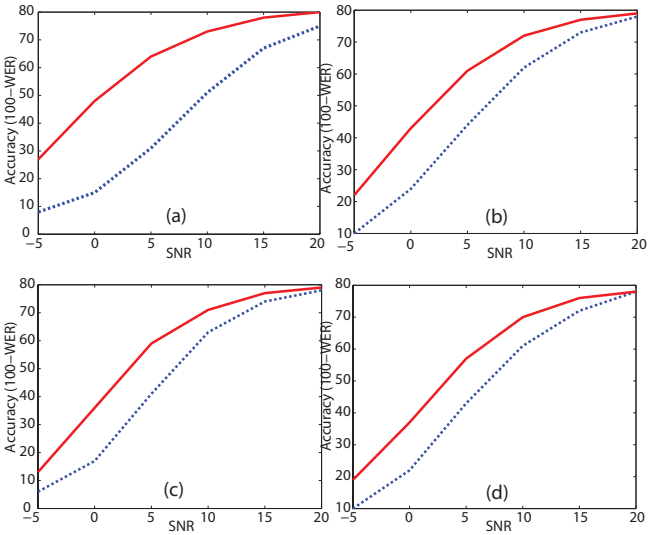

Fig. 1. Recognition performance on speech corrupted by a. classical, b. jazz, c. latin, d. world music. In each figure, the lower dotted curve is the performance on uncompensated speech and the upper curve is the performance on enhanced speech.

4.2. Speaker and music style dependent bases

We ran a number of different experiments. In the first set of experiments, we corrupted the test speech with each of four music types: classical, jazz, latin, and world to a number of different SNRs, namely -5dB, 0dB, 10dB, 15dB and 20dB. NMF-based separation is usually assumed to require detailed knowledge of the corrupting noise and the speaker. This is often not such an unrealistic assumption – the identity of the speaker will often be known, and the bases for music can be learned from music-only segments that have been detected by a voice-activity detector. For this test we used the training sets of recordings from each speaker to learn speaker-dependent bases for each speaker. For each of the music types we also learned music-specific bases. 3000 bases were obtained for each music type and speaker. In all experiments, the signals were analyzed using 60ms windows with a 15ms frame shift between windows to compute spectrograms.

Figure 1 shows the performance on speech corrupted by each of the four categories of music. We observe, firstly, that significant improvements in recognition accuracy are obtained on speech corrupted by all music types. The most improvement, however, is obtained on speech corrupted by classical music.

4.3. Multi-speaker and multi-style music bases

This first experiment makes some pretty stringent assumptions. It assumes that the type of music affecting the signal and the identity of the speaker are both known. In the next two experiments we relaxed both assumptions. Speech was corrupted with randomly selected segments of music from any of the music types from the RWC Genre database. In the first experiment the identity of the corrupting music was assumed to be unknown. A total of 3000 bases were drawn randomly from *l all* music types and used for separation. Although this is not an open set of music types, it is still a fairly large closed set since the music segments are diverse in their variety. In the second experiment it was assumed that the

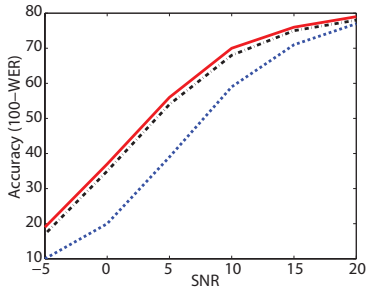

Fig. 2. Recognition of speech corrupted by assorted music. The lowest curve shows performance on uncompensated speech. The central black dash-dotted line is obtained with mixed music bases and speaker-specific speech bases. The top red line is the performance with mixed music bases and multi-speaker speech bases.

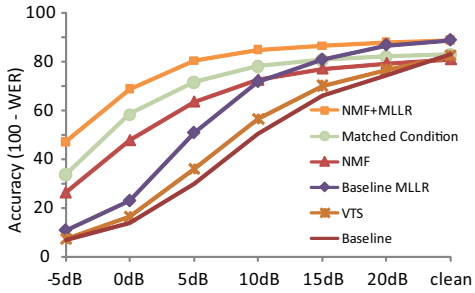

Fig. 3. Recognition performance on speech corrupted by classical music. Results shown are i) basline on uncompensated speech ii) on speech compensated by VTS [3], iii) with baseline models adapted to the corrupted data using MLLR, iv) on NMF-enhanced speech, v) with a "matched" recognizer, and vi) after adaptation of the models using hypotheses obtained from NMF-enhanced speech, on NMF-enhanced speech.

identity of the speaker too was unknown. A common set of 6000 bases drawn from all speakers was used for separation. Again, although this is not an open set of speakers, the number of speakers (83) is very large. Figure 2 shows the performance obtained in both experiments. The algorithm not only holds up when the identity of the music and speaker are unknown, but the performance obtained when the speaker identity was not known is actually slightly *higher* than that obtained when the identity is known.

4.4. Recognizer adaptation

Speech recognition systems frequently employ adaptation techniques to improve recognition. It is often unclear if compensation methods combine well with adaptation, and if they do what the upper bound on performance might be. Figure 3 shows the performance obtained with maximum likelihood linear regression (MLLR) adaptation. For this experiment we chose the speech data corrupted by music and speaker-specific speech bases were employed. Adaptation was performed by speaker.

Not only does adaptation improve performance greatly, the final performance is better than that obtained with a *matched* recognizer that was trained on speech corrupted by exactly the same music and music level as the test speech. This is the best performance we have obtained to date on speech corrupted by music.

5. CONCLUSIONS

We have shown that NMF-based compensation of speech corrupted by music can result in large improvements in recognition accuracy. We have also shown that although the compensation requires bases drawn from the music and speech, it functions very well even when the identity of the music or speaker are unknown. Interestingly, we observe that large improvements are obtained in recognition accuracy even when when perceptual improvements in the background music level in the signal are not as high.

Various direct enhancements to NMF, including the enforcement of temporal continuity constraints, can improve performance greatly. In addition, NMF provides an instantaneous characterization of the distribution of music, through the weights assigned to the bases. This can in fact be used directly to adapt the models in the recognizer for improved recognition. This and other techniques remain topics for future work.

6. REFERENCES

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. on ASSP*, 1979.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error log-spectral amplitude estimator," *IEEE Trans. on ASSP*, pp. 443–445, 1985.

[3] P. J. Moreno, B. Raj, and R. M. Stern, "A Vector Taylor Series Approach for Environment-Independent Speech Recognition," *Proc. ICASSP*, 1996.

[4] D. Yu, L. Deng, J. Droppo, J. Wu, Y. Gong, and A. Acero, "A Minimum-Mean-Square-Error Noise reduction Algorithm on Mel-Frequency Cepstra for Robust Speech Recognition," *Proc. ICASSP*, 2008.

[5] J. Droppo and A. Acero, "Noise Robust Speech Recognition with a Switching Linear Dynamic Model," *Proc. ICASSP*, 2004.

[6] B. Raj, R. Singh, and R. M. Stern, "On Tracking Noise with Linear Dynamical System Models," *Proc. ICASSP*, 2004.

[7] A. P. Varga and R. K. Moore, "Hidden Markov Model decomposition of speech and noise," *Proc. ICASSP*, 1990.

[8] J. R. Hershey, S. J. Rennie, O. P. A., and T. T. Kristjansson, "Super-human multi-talker speech recognition: A graphical modeling approach," *Computer Speech and Language*, 2010.

[9] T. Virtanen, "Monaural Sound Source Separation by Non-Negative Matrix Factorization with Temporal Continuity and Sparseness Criteria," *IEEE Trans. on ASLP*, vol. 15, 2007.

[10] M. V. Shashanka, B. Raj, and P. Smaragdis, "Probabilistic Latent Variable Models as Non-Negative Factorizations," *Computational Intelligence and Neuroscience*, May 2008.

[11] P. Smaragdis, R. Shashanka, and B. Raj, "A Sparse Non-Parameteric Approach for Single Channel Separation of Known Sounds," *Proc. NIPS*, 2009.

[12] J. F. Gammecke and T. Virtanen, "Noise-robust exemplar-based connected digit recognition," *Proc. ICASSP*, 2010.

[13] A. Cichocki, "Csiszar Divergences for Non-negative Matrix Factorization: Family of New Algorithms," *ICA and BSS*, vol. 3889/2006, pp. 32–39, 2006.

[14] M. Goto, "Development of the RWC music database," in *the 18th International Congress on Acoustics*, 2004.